# The Weighted Set Cover Problem

- We are given a set $S$ of size $n$, and a family $\mathcal{F}$ of subsets of $S$
  - Each set $T$ in $\mathcal{F}$ has an associated nonnegative cost $c(T)$
  - A set cover is a subset $\mathcal{C}$ of $\mathcal{F}$ such that $\cup_{T \in \mathcal{C}} T = S$
  - The cost of a set cover $\mathcal{C}$ is $\sum_{T \in \mathcal{C}} c(T)$
  - We seek a minimum-cost set cover
  - We assume that $\mathcal{F}$ is itself a set cover, so a solution is guaranteed to exist

# NP-Hardness of Set Cover

- The vertex cover problem corresponds to a special case of the set cover problem
  - Let $G = (V, E)$ be an instance of the (unweighted) vertex cover problem
  - The set $S$ of elements to be covered is $E$
  - The family $\mathcal{F}$ of subsets of $S$ includes one set $T_v$ for each vertex in $V$, namely, the set of all edges incident on $v$
  - Every set in $\mathcal{F}$ has unit cost
- Thus the (unweighted) set cover problem is NP-hard

# A Greedy Algorithm

- We initialize $\mathcal{C}$ to $\emptyset$, and we repeatedly apply a greedy rule to determine a set in $\mathcal{F}$ to add to $\mathcal{C}$, terminating when $\mathcal{C}$ is a set cover
- The greedy rule selects the "best bang for the buck" set
  - Let $S'$ denote the uncovered elements $S \setminus (\cup_{T \in \mathcal{C}} T)$
  - We select a set $T$ in $\mathcal{F}$ minimizing $c(T)/|S' \cap T|$

# A Price-Based Analysis of the Greedy Algorithm

- For the purposes of analysis, it is useful to assign a price $p(e)$ to each element $e$ of $S$, as follows
  - Let $T$ be the first set selected by the algorithm that includes $e$
  - We set $p(e)$ to the ratio $c(T)/|S' \cap T|$ in the iteration that selected $T$
  - Thus the sum of the prices determined in this iteration is $\sum_{e \in S' \cap T} p(e) = c(T)$
- Upon termination, the cost of the set cover $\mathcal{C}$ is equal to $\sum_{e \in S} p(e)$

# An Upper Bound for the Greedy Algorithm

- Let $e_i$ be the $i$th element of $S$ covered by the greedy algorithm, $1 \leq i \leq n$, breaking ties arbitrarily
- Let $C^*$ denote the cost of a minimum-cost set cover
- Lemma 1: $p(e_i) \leq C^*/(n - i + 1)$
  - In the iteration in which $e_i$ is covered, we have $|S'| \geq n - i + 1$
  - The elements in $S'$ can be covered at a cost of at most $C^*$
  - Thus $c(T)/|S' \cap T| \leq C^*/(n - i + 1)$ for the selected set $T$

# An Upper Bound for the Greedy Algorithm (cont'd)

- Lemma 1 implies that the total cost of the set cover produced by the greedy algorithm is at most

$$\sum_{1 \le i \le n} \frac{C^*}{n - i + 1} = C^* \sum_{1 \le i \le n} \frac{1}{i} = C^* H_n$$

- Thus the greedy algorithm achieves an approximation ratio of $H_n \sim \ln n$

# A Bad Example for the Greedy Algorithm

- Let $S = \{e_1, \ldots, e_n\}$
- Let $\mathcal{F} = \{T_1, \ldots, T_{n+1}\}$ where the sets $T_i$ are defined as follows
  - For any integer $i$ such that $1 \leq i \leq n$, we have $T_i = \{e_i\}$ and $c(T_i) = \frac{1}{n-i+1}$
  - We have $T_{n+1} = S$ and $c(T_{n+1}) = 1 + \varepsilon$ for an arbitrarily small $\varepsilon > 0$
- How does the greedy algorithm behave on this instance?

# A Bad Example for the Greedy Algorithm (cont'd)

- In the $i$th round, the greedy algorithm selects $T_i$ because it has cost ratio $1/(n-i+1)$
  - The sets $T_j$ with $1 \leq j < i$ have already been selected and thus have infinite cost ratio
  - For any integer $j$ such that $i \leq j \leq n$, the set $T_j$ has cost ratio $1/(n-j+1)$
  - The set $T_{n+1}$ has cost ratio $(1+\varepsilon)/(n-i+1)$
- The greedy set cover has cost

$$\sum_{1 \leq i \leq n} \frac{1}{n-i+1} = H_n \sim \ln n$$

- For $n \geq 2$, the optimal set cover has cost $1+\varepsilon$

# A Bad Example for the Unweighted Case

- Even if we require $c(T) = 1$ for all sets $T$ in $\mathcal{F}$, the worst-case approximation ratio achieved by the greedy algorithm is $\Omega(\log n)$
- Let $S = A \cup B$ where $A = \{a_1, \ldots, a_{n/2}\}$ and $B = \{b_1, \ldots, b_{n/2}\}$ are disjoint, and $n = 2(2^k - 1)$ for some integer $k > 0$
  - Let $A_0$ denote $\{a_1\}$, let $A_1$ denote $\{a_2, a_3\}$, let $A_2$ denote $\{a_4, a_5, a_6, a_7\}$, et cetera
  - Thus the sets $A_0, \ldots, A_{k-1}$ form a partition of $A$
  - Similarly, we partition $B$ into sets $B_0, \ldots, B_{k-1}$
- Let $\mathcal{F} = \{T_0, \ldots, T_{k-1}, A, B\}$ where $T_i = A_i \cup B_i$ for $0 \leq i < k$

# A Bad Example for the Unweighted Case (cont'd)

- In the first iteration, the greedy algorithm selects $T_{k-1}$ since it is the largest of the $T_i$'s and $|T_{k-1}| = 2 \cdot 2^{k-1} = 2^k$ while $|A| = |B| = 2^k - 1$
- In the second iteration, the greedy algorithm selects $T_{k-2}$ since $|T_{k-2} \cap S'| = 2 \cdot 2^{k-2} = 2^{k-1}$ while $|A \cap S'| = |B \cap S'| = 2^{k-1} - 1$
- This continues for $k$ iterations, until the greedy algorithm has selected all of the $T_i$'s
- There is a set cover $\{A, B\}$ of cardinality 2
- Thus the worst-case approximation ratio achieved by the greedy algorithm is $k/2 = \Omega(\log n)$

# Inapproximability of Set Cover

- Even in the unweighted case, it is known that no polynomial-time algorithm achieves a $(1 - o(1)) \ln n$ approximation ratio for set cover unless $\mathrm{P} = \mathrm{NP}$
  - The proof of this claim is beyond the scope of this course
- Thus, assuming $\mathrm{P} \neq \mathrm{NP}$, the greedy algorithm that we have presented provides essentially the best possible polynomial-time approximation guarantee
- Many hardness of approximation results in the literature are based on approximation-preserving reductions from set cover

# Approximating Set Cover via LP Duality

- As you might guess, our price-based analysis of the greedy set cover algorithm has a connection to LP duality
- In what follows, we consider two ways to use LP duality to obtain an approximation algorithm for the weighted set cover problem
  - One of these two approaches corresponds to the greedy algorithm presented earlier

# A 0-1 ILP Formulation of Weighted Set Cover

- We have a 0-1 variable $x_T$ for each set $T$ in $\mathcal{F}$
- For each element $e$ in $S$, we have a "covering constraint"

$$\sum_{T \in \mathcal{F} : e \in T} x_T \geq 1$$

- The objective is to minimize $\sum_{T \in \mathcal{F}} c(T) x_T$
- In the corresponding LP relaxation, for each $T$ in $\mathcal{F}$ we relax the constraint $x_T \in \{0, 1\}$ to $x_T \geq 0$
  - We refer to the LP relaxation as the primal LP

# The Dual of the LP Relaxation

- We can mechanically form the dual of the primal LP
- We have a nonnegative variable $y_e$ for each element $e$ in $S$
- For each set $T$ in $\mathcal{F}$, we have the "packing constraint"

$$\sum_{e \in T} y_e \leq c(T)$$

- The objective is to maximize $\sum_{e \in S} y_e$

# An Algorithm Based on the Primal-Dual Schema

- Here we proceed as in the development of the price-based approximation algorithm for vertex cover presented in the previous lecture
  - We maintain a feasible solution $y$ that is initialized to the all-zeros vector
  - The corresponding 0-1 solution, which may be infeasible, sets $x_T = 1$ if and only if the packing constraint corresponding to $T$ is tight
  - While $x$ is infeasible, we identify an element $e$ of $S$ for which the covering constraint is violated, and we raise $y_e$ until some packing constraint involving $y_e$ becomes tight

# An Upper Bound for the Primal-Dual Algorithm

- Let $k$ denote the maximum, over all elements $e$ in $S$, of $|\{T \in \mathcal{F} \mid e \in T\}|$
  - Remark: In the special case of vertex cover, we have $k = 2$
- The primal-dual algorithm achieves an approximation ratio of $k$ for weighted set cover
  - Let $\mathcal{C}$ be the set cover computed by the algorithm
  - The cost of $\mathcal{C}$ equals $\sum_{T \in \mathcal{C}} c(T)$ which is equal to $\sum_{T \in \mathcal{C}} \sum_{e \in T} y_e \le k \sum_{e \in S} y_e$
  - The lemma follows since the dual solution $y$ is feasible and has objective function value $\sum_{e \in S} y_e$

# A Bad Example for the Primal-Dual Algorithm

- Consider an instance with $S = \{e_1, \ldots, e_n\}$ where $n \geq 3$
- The family $\mathcal{F} = \{T_1, \ldots, T_{n-1}\}$ of subsets of $S$, where the $T_i$'s are defined as follows
  - $T_1 = S$ and $c(T_1) = 1 + \varepsilon$ for an arbitrarily small $\varepsilon > 0$
  - For any integer $i$ such that $2 \leq i < n$, we have $T_i = \{e_2, e_{i+1}\}$ and $c(T_i) = 1$
- If the primal-dual algorithm begins by raising $y_{e_2}$ to 1, then it produces the set cover $\mathcal{F} \setminus \{T_1\}$ with cost $n - 1$
- The set cover $\{T_1\}$ has cost $1 + \varepsilon$

# The "Dual Fitting" Method

- In the dual fitting method (as applied to a minimization problem), we maintain primal-dual solutions satisfying the following conditions
  - The primal solution is integral and is feasible upon termination
  - The objective function value of the primal solution is at most the objective function value of the dual solution
  - The dual solution is nonnegative but need not be feasible
  - If we divide the dual solution by some factor $\alpha > 1$, it becomes feasible
- Next, we argue that the greedy algorithm presented earlier corresponds to an application of the dual fitting method with $\alpha$ set to $H_n$

# Revisiting the Greedy Algorithm

- Upon termination, let $y_e$ denote $p(e)/H_n$ for each $e$ in $S$
- Lemma 3: The dual solution $y$ is feasible
  - Let $T$ be a set in $\mathcal{F}$
  - The $i$th item covered in $T$ has price at most $c(T)/(|T|-i+1)$
  - Thus

$$\sum_{e \in T} y_e \leq \frac{1}{H_n} \sum_{1 \leq i \leq |T|} \frac{c(T)}{|T|-i+1} = \frac{H_{|T|}}{H_n} \cdot c(T) \leq c(T)$$

# Revisiting the Greedy Algorithm (cont'd)

- The greedy algorithm maintains the invariant that the sum of the prices is equal to the cost of the selected sets
- Thus, upon termination, the cost of the set cover is equal to $\sum_{e \in S} p(e)$
- By Lemma 3 and the weak duality theorem, the optimal objective function value for the primal is at least $(1/H_n) \sum_{e \in S} p(e)$
- Thus the approximation ratio achieved by the greedy algorithm is at most $H_n \sim \ln n$

# The Integrality Gap of the Set Cover LP

- We will prove that the integrality gap of (unweighted) set cover is $\Omega(\log n)$, where $n$ denotes the size of the set to be covered
- We will construct an infinite family of set cover instances parameterized by a positive integer $k$
- For any $k$, the associated set cover instance is defined in terms of the vector space $\mathbb{F}_2^k$
- We begin by reviewing some basic facts about $\mathbb{F}_2^k$

# The Vector Space $\mathbb{F}_2^k$

- The vector space $\mathbb{F}_2^k$ has $2^k$ elements
  - Each element is a 0-1 vector of length $k$
  - Addition in $\mathbb{F}_2$ corresponds to $\oplus$
  - Multiplication in $\mathbb{F}_2$ corresponds to $\wedge$
  - The inner product $\langle u, v \rangle$ of two vectors $u$ and $v$ in $\mathbb{F}_2^k$ is defined in the usual manner, except addition and multiplication are performed in $\mathbb{F}_2$

# The Set Cover Instance $I_k$

- Let $V$ denote $\mathbb{F}_2^k$ and let $V^*$ denote $V$ minus the all-zeros vector
- For any $u$ in $V$, let $T_u$ denote

$$\{v \in V \mid \langle u, v \rangle = 1\} = \{v \in V^* \mid \langle u, v \rangle = 1\}$$

- We define the set of elements to be covered as $V^*$ and the family $\mathcal{F}$ of subsets of $V^*$ as $\{T_u \mid u \in V\}$
  - Thus $|V^*| = 2^k - 1$ and $|\mathcal{F}| = 2^k$

# A Key Claim

- Lemma 4: Each vector in $V^*$ belongs to exactly half of the sets in $\mathcal{F}$
  - Let $v$ be an arbitrary vector in $V^*$
  - Let $i$ be an index such that $v_i \neq 0$; such an index exists since $v$ is not the all-zeros vector
  - Let $u$ be a uniformly random vector in $V$
  - We have $\langle u, v \rangle = \langle u_{-i}, v_{-i} \rangle + u_i$
    - Here $u_{-i}$ (resp., $v_{-i}$) denotes the vector $u$ (resp., $v$) with component $i$ removed
  - By deferring the random choice of $u_i$ until after $u_{-i}$ has been chosen, it is easy to see that $\Pr(\langle u, v \rangle = 1) = \frac{1}{2}$

# A Good Fractional Solution

- The relaxed set cover LP has a variable $x_T$ for each set $T$ in $\mathcal{F}$
- We claim that by setting each variable $x_T$ to $\frac{2}{|\mathcal{F}|}$, we obtain a feasible solution
  - Fix a vector $v$ in $V^*$
  - By Lemma 4, we have $\sum_{T \in \mathcal{F} : v \in T} x_T = \frac{|\mathcal{F}|}{2} \cdot \frac{2}{|\mathcal{F}|} = 1$
- This feasible solution has an objective function value of 2
  - We have $\sum_{T \in \mathcal{F}} x_T = |\mathcal{F}| \cdot \frac{2}{\mathcal{F}} = 2$

# A Lower Bound for any Integral Solution

- Let $\mathcal{C} = \{T_{u_1}, \ldots, T_{u_\ell}\}$ be a set cover
- For any $i$, $0 \leq i \leq \ell$, let $V_i$ denote $V \setminus (\cup_{1 \leq j \leq i} T_{u_j})$
- Thus $V_0 = V$ and for $1 \leq i \leq \ell$, $V_i$ is the subspace of all vectors $v$ in $V_{i-1}$ such that $\langle u_i, v \rangle = 0$
- Thus the dimension of $V_i$ is at most one less than the dimension of $V_{i-1}$ for $1 \leq i \leq \ell$
- Since $V$ has dimension $k$, $V_\ell$ has dimension at least $k - \ell$
- Since $\mathcal{C}$ is a set cover, $V_\ell \cap V^* = \emptyset$ and hence the dimension of $V_\ell$ is zero
- We conclude that $\ell \geq k$

# A Lower Bound for the Integrality Gap

- Instance $I_k$ admits a fractional solution with objective function value 2
- Any integral solution has objective function value at least $k$
- The cardinality of the set $V^*$ to be covered is $2^k - 1$
- Thus the integrality gap is at least $\frac{k}{2}$
- Letting $n$ denote $2^k - 1$, we find that the integrality gap is $\Omega(\log n)$